Ability of the misfolding avoidance hypothesis to explain protein evolutionary rates in the light of empirical evidence

Dinara R. Usmanova[1*], Germán Plata[1,2*], Dennis Vitkup[1,3†]

[1] Department of Systems Biology, Columbia University, 1130 Saint Nicholas Ave., New York, NY 10032, USA

[2] Elanco Animal Health, 2500 Innovation Way, Greenfield, IN 46140, USA

[3] Department of Biomedical Informatics, Columbia University, 622 West 168th street, New York, NY 10032, USA

[*] These authors contributed equally

[†] Correspondence to DV (dv2121@columbia.edu)

ABSTRACT

In this study we further discuss the relationship between the rate of protein evolution and the misfolding avoidance hypothesis (MAH), and clarify the use of experimental data to test the hypothesis. We demonstrate, in *Escherichia coli* and *Homo sapiens*, the lack of a substantial negative correlation between protein evolutionary rates and Gibbs free energies of unfolding, a direct measure of protein stability. We also analyze genome-scale data describing protein aggregation and interaction propensities, which are likely optimized to alleviate deleterious effects associated with toxic protein misfolding and misinteractions. Our results demonstrate that the propensity of proteins to aggregate, the fraction of charged amino acids, and protein stickiness do correlate with protein abundances. Nevertheless, across multiple organisms and datasets we did not observe a substantial effect of aggregation-related properties on the variability of protein evolutionary rates. Therefore, diverse empirical data support the conclusion that the MAH is unlikely to play a major role in explaining the strong negative correlation between protein abundance and the protein molecular clock.

INTRODUCTION

Protein evolutionary rates vary by orders of magnitude, but the mechanisms underlying this variability are currently unknown [1]. Although protein expression was shown to be the strongest predictor of protein evolutionary rates across species [2, 3], the causes of the anticorrelation between expression and protein evolutionary rate are not understood [4]. The popular misfolding avoidance hypothesis (MAH) posits that the sequences of highly abundant proteins evolve slowly *primarily* due to increased selection against misfolded protein toxicity [5-7]. The recent availability of genome-wide experimental data on protein stability has reinvigorated the debate about the model of protein evolution based on the MAH [8, 9].

We read with interest the recent manuscript "Protein melting temperature cannot fully assess whether protein folding free energy underlies the universal abundance-evolutionary rate correlation seen in proteins" by Rostam Razban [9]. This paper is related to our previous analyses, i.e. Plata *et al.* (2010) [10] and especially Plata and Vitkup (2018) [8]. Razban's manuscript discusses our results showing a lack of empirical support for the MAH based on the genome-wide protein melting temperature (T_m) data obtained by Leuenberger *et al.* [11]. To avoid potential misunderstanding in the field, in this paper we address inaccuracies in the characterization of our previous work, and comment more broadly on proper experimental tests of the MAH. We also analyze available Gibbs unfolding free energies (ΔG) for *E. coli* and *H. sapiens* proteins [12], and several proteome-wide datasets characterizing protein aggregation and misinteraction propensities in several species [13-17].

The MAH can be tested by investigating the two key predictions of the hypothesis: (i) protein abundance positively correlates with protein stability [4, 6], and (ii) protein stability substantially affects the variation of evolutionary rates across cellular proteins. As we demonstrated previously, a careful re-analysis of the proteome-wide T_m measurements obtained by Leuenberger *et al.* [11] shows no support for the MAH in *E. coli* and three other investigated species [8]. Razban's manuscript states that our analysis provides support for the MAH in *E. coli* due to a correlation between protein abundances and melting temperatures. This claim is not correct and, we believe, exemplifies a common and unfortunate confusion. As we specifically discussed [8, 10], the MAH cannot be validated simply by demonstrating a weak correlation between protein abundance and stability, i.e. the relationship (i) above. The MAH, at its core, is not about the stability of highly expressed proteins, but about a major effect of protein stability on the level of sequence constraints across cellular proteins. Thus, it is essential to investigate whether protein stability accounts for any substantial fraction of the variance of evolutionary rates across proteins. Such analysis, based on either the direct correlation between the rate of protein evolution and protein stability, or partial correlation between the rate of evolution and abundance, while controlling for stability, showed no support for a major role of MAH in any considered organism [8].

RESULTS

The central message of Razban's analysis is that the absence of the expected relationship between T_m and protein abundance may be due to an imperfect correlation between T_m and ΔG, as well as possible experimental noise in T_m measurements [9]. To address the issue of the correlation between protein melting temperature and stability, we analyzed available data describing protein unfolding Gibbs free energies, ΔG. Empirical ΔG values have now been obtained for a substantial number of proteins in *E. coli* and *H. sapiens*, and are available in the ProTherm database [12]. Notably, for the subsets of proteins with available unfolding free energies we can robustly reproduce a significant anticorrelation between evolutionary rates

and mRNA expression (**Figure 1**, blue; **Supplementary Table 1**; Spearman's $r = -0.56,\ p < 2 \cdot 10^{-3}$, for *E. coli*; $r = -0.54,\ p < 2 \cdot 10^{-4}$, for *H. sapiens*). Thus, the mechanisms that make highly expressed proteins evolve slower are likely to be reflected in the properties of these proteins. Nevertheless, we didn't find a negative correlation between ΔG and evolutionary rates (**Figure 1**, red), or positive correlations between ΔG and either protein abundances or mRNA expression levels (**Figure 1**, light and dark grey, respectively) predicted by MAH. Therefore, in agreement with the conclusions based on T_m from Leuenberger *et. al* [8], the empirical ΔG data also do not provide any support for a major role of the MAH in explaining the variation of evolutionary rates across proteins.

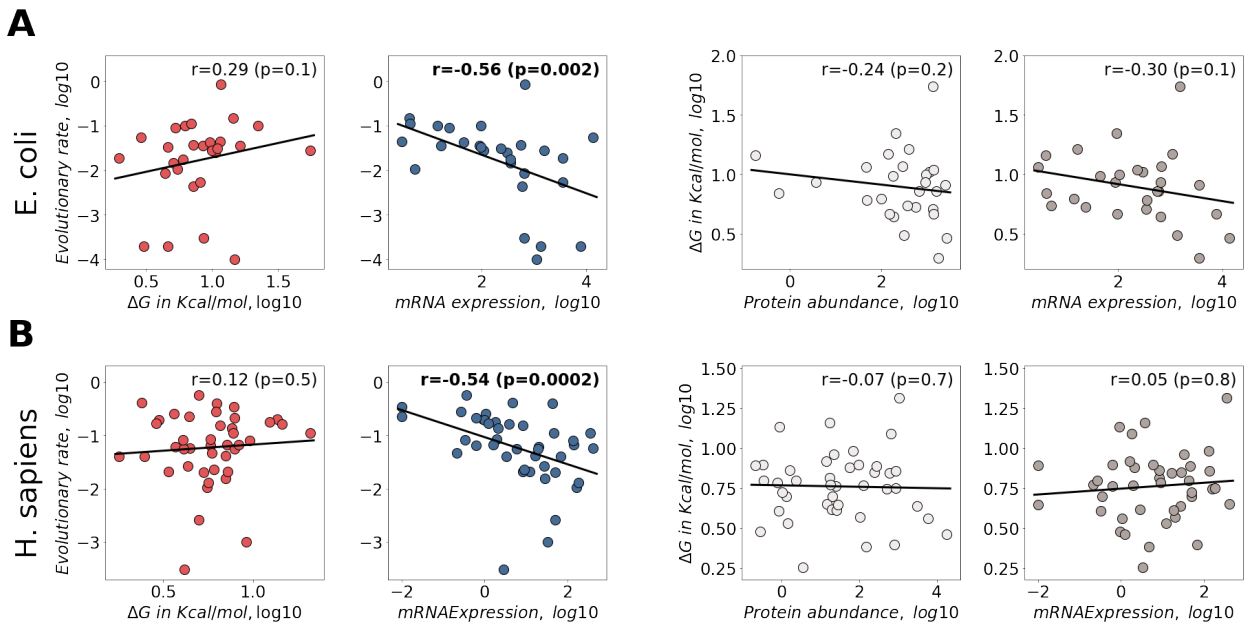

Figure 1. Correlations between experimentally measured ΔG values, protein abundances, mRNA expression levels, and protein evolutionary rates.

A. *E. coli* (n=28) and **B.** *H. sapiens* (n=47). The correlations between evolutionary rates and unfolding Gibbs free energies, ΔG, are shown in the first figure column (red). The correlations between protein evolutionary rates and mRNA expression are shown in the second column (blue). The correlations between ΔG and protein abundances are shown in the third column (light grey), and the correlations between ΔG and mRNA expression are shown in the forth column (dark grey). Solid lines represent the least square regressions fitted to the data. Spearman's correlation coefficients and corresponding p-values are shown, significant correlations are highlighted in bold.

In our view, the main caveat with Leuenberger *et al.* dataset, in the context of testing the MAH, is not a poor correlation between T_m and ΔG; we previously demonstrated a significant correlation (Pearson's $r = 0.75,\ p < 10^{-20}$; Spearman's $r = 0.64,\ p < 10^{-20}$) between these two characteristics of protein stability across measurements performed by the same research group [8]. Instead, intrinsic protein stability may simply not serve as a good proxy for protein aggregation propensity, which is likely to mediate misfolding toxicity. The protein melting temperatures obtained by Leuenberger *et al.* [11] are based on data from limited proteolysis (LiP), which increases due to local protein unfolding triggered by higher temperatures; below, we use the term T_m^{LiP} for these melting temperature measurements. An alternative method, developed by Savitski *et al.* [13], uses protein aggregation (Agg) as a proxy of unfolding. This

method estimates melting temperatures by quantifying protein concentrations in soluble cellular fractions as a function of increasing temperature; we use the term T_m^{Agg} for these melting temperature measurements. It is likely that T_m^{Agg} is more relevant for testing the MAH than either T_m^{LiP} or ΔG, because it reflects the relative propensity of proteins to aggregate, and therefore to ultimately cause misfolding toxicity.

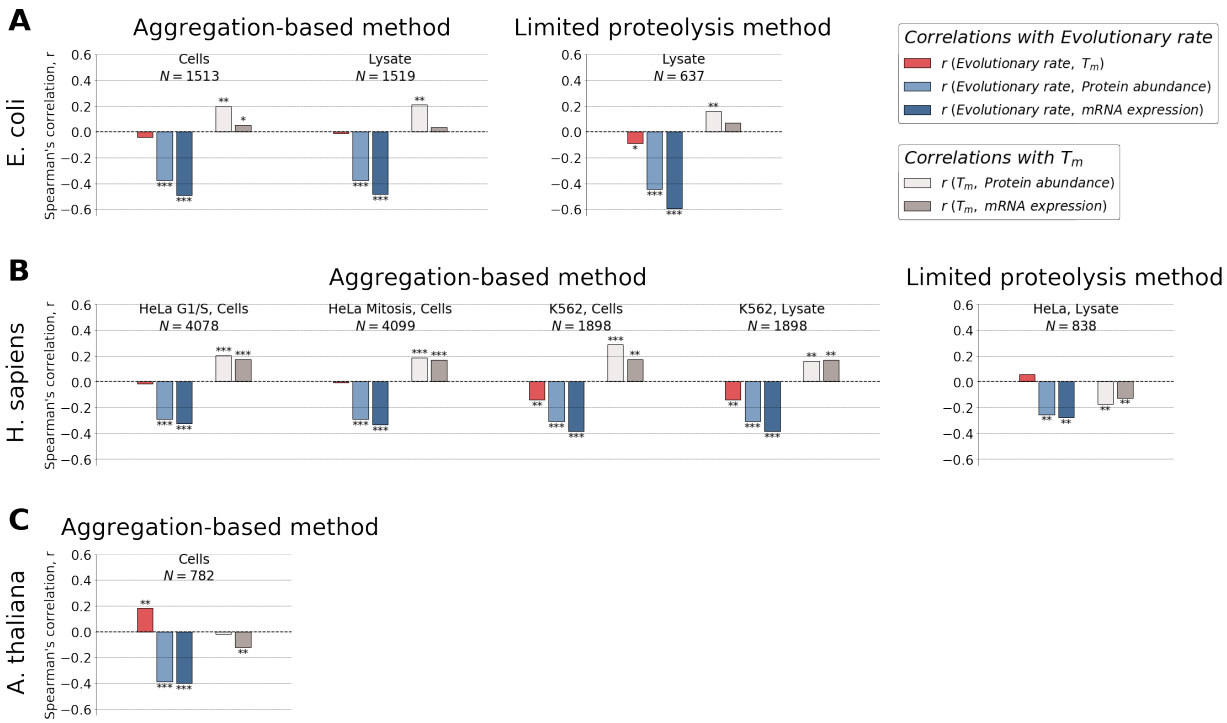

Figure 2. Correlations between genome-wide melting temperatures, protein abundances, mRNA expression levels, and protein evolutionary rates.

A, *E. coli*, **B**, *H. sapiens* and **C**, *A. thaliana.* Bar plots show the values of the Spearman's correlation coefficients between evolutionary rates and melting temperatures (red), between evolutionary rates and protein abundances (light blue), between evolutionary rates and mRNA expression (dark blue), between melting temperatures and protein abundances (light grey), and between melting temperatures and mRNA expression (dark grey). Different methodologies used to measure T_m and different sample types are indicated above figure panels for *E. coli* and *H. sapiens*. Numbers of proteins in the analyzed datasets are shown, in each dataset we kept only proteins for which all four parameters are known. Asterisks above and below bars indicate significance level: * for p-value < 0.05, ** for p-value < 0.001, *** for p-value < 10^{-25}.

To specifically analyze the potential effects of protein aggregation on protein evolution we used the T_m^{Agg} data for ~1500 *E. coli* proteins based on measurements performed in cells and natural cellular lysates [15]. Interestingly, we found that in both datasets T_m^{Agg} significantly correlated with protein abundances (**Figure 2A**, light grey; **Supplementary Table 2**; Spearman's $r = 0.20$, $p < 10^{-14}$, for cells; $r = 0.21$, $p < 10^{-15}$, for cell lysates). However, we observed no significant correlations between T_m^{Agg} and evolutionary rates (**Figure 2A**, red). T_m^{Agg} were also independently measured in two different *H. sapiens* cell lines: HeLa cells [14], where the T_m^{Agg} of ~4000 proteins were obtained for intact cells in

different cell-cycle stages (G1/S transition and mitosis), and K562 chronic myeloid leukemia cells [13], where the T_m^{Agg} of ~2000 proteins were obtained for intact cells and cellular lysates. Analyzing these measurements, we found that across all the human datasets T_m^{Agg} also positively correlated with protein abundances (**Figure 2B**, light grey; Spearman's $r = 0.20, p < 10^{-38}$; $r = 0.19, p < 10^{-32}$; $r = 0.29, p < 10^{-36}$; and $r = 0.16, p < 10^{-11}$). For the K562 datasets, T_m^{Agg} values were also negatively correlated with protein evolutionary rates (**Figure 2B**, red; Spearman's $r = -0.14$, $p < 10^{-9}$, for both cells and lysate). Finally, T_m^{Agg} data were also obtained for ~800 proteins in Arabidopsis [16]. In this case, we did not find any correlation of T_m^{Agg} with protein abundances (**Figure 2C**, light grey) and the correlation with evolutionary rates was significant, but positive (**Figure 2C**, red; Spearman's $r = 0.18$, $p < 10^{-6}$).

What fraction of the molecular clock variance is explained by the observed correlations with T_m^{Agg}? According to the MAH, avoiding aggregation toxicity is a major driver of the variability in protein evolutionary rates. However, our analyses demonstrate that the anticorrelation between T_m^{Agg} and evolutionary rates (**Figure 2**, red) is significant only in two (out of seven) datasets, and even in these two T_m^{Agg} explain only ~2% of the variance in evolutionary rates across proteins. For comparison, mRNA expression explains about an order of magnitude higher fraction of the evolutionary rate variance (**Figure 2**, dark blue), i.e. ~15% for the same subset of proteins. Furthermore, using partial correlation analysis, we found that the anticorrelations between mRNA expression and evolutionary rates do not substantially decrease after controlling for T_m^{Agg} (for example, Spearman's $r_{Ev.Rate-Exp.} = -0.38$ and corresponding Spearman's partial $r_{Ev.Rate-Exp|T_m^{Agg}} = -0.37$ for *H. sapiens* K562 cells, the dataset with the strongest effects of T_m^{Agg}).

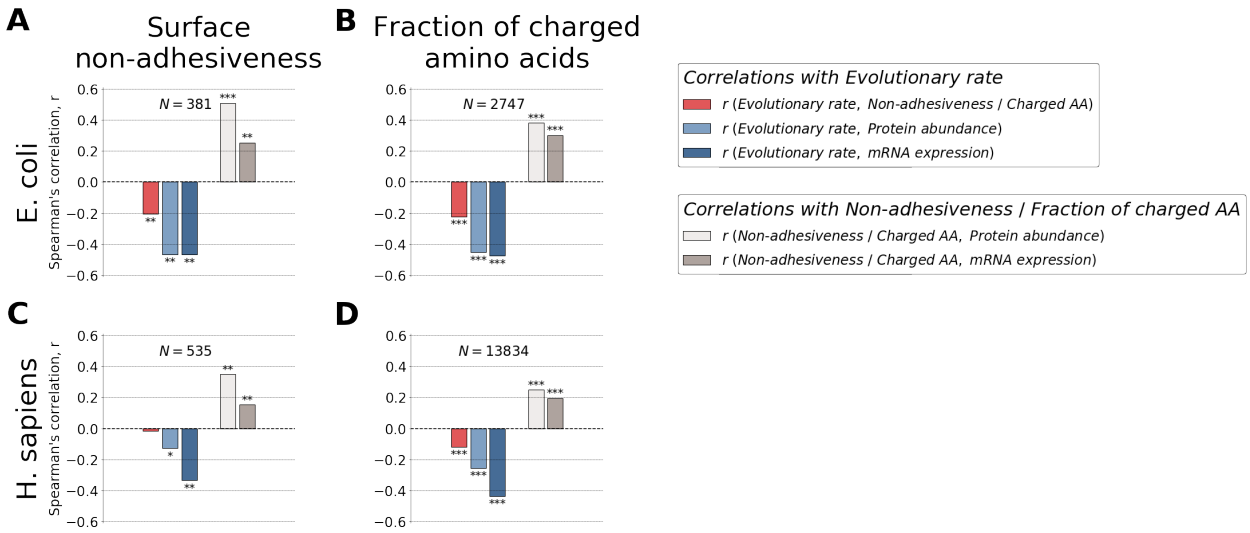

Figure 3. Correlations between protein surface non-adhesiveness, the fraction of charged amino acids, protein abundances, mRNA expression levels, and protein evolutionary rates.

A, **B**, *E. coli*, **C**, **D**, *H. sapiens*. Bar plots in **A** and **C** show the values of Spearman's correlation coefficients between evolutionary rates and protein surface non-adhesiveness (red), between protein surface non-adhesiveness and protein abundance (light grey), and between protein surface non-adhesiveness and mRNA expression (dark grey). Bar plots in **B** and **D** show the values of Spearman's correlation coefficients between evolutionary rates and the fraction of charged amino acids (red), between the fraction of charged amino acids and protein abundance (light grey), and

between the fraction of charged amino acids and mRNA expression (dark grey). In all panels the values of Spearman's correlation coefficients between evolutionary rates and protein abundances (light blue), between evolutionary rate sand mRNA expression (dark blue) are also shown. Numbers of proteins in the analyzed datasets are indicated, in each dataset we kept only proteins for which all four parameters are known. Asterisks above and below bars represent significance level: * for p-value < 0.05, ** for p-value < 0.001, *** for p-value < 10^{-25}.

To put the observed correlations with T_m^{Agg} into perspective, we note that multiple other protein properties, such as the fraction of charged amino acids [10], protein solubility [10], surface stickiness [17], and the number of protein-protein interaction partners [18], have been shown to correlate with protein abundance. Changes in these properties with protein abundance detected in *E. coli* [10, 17], *S. cerevisiae* [17, 18] and *H. sapiens* [17], likely help to alleviate deleterious effects of non-functional interactions and binding. For example, protein surface non-adhesiveness or the fraction of charged amino acids correlate positively, and with similar strength as T_m^{Agg}, with protein abundances (**Figure 3**, light grey), and negatively with evolutionary rates (**Figure 3**, red, **Supplementary Table 3**, **Supplementary Table 4**). However, the ability of all these protein characteristics to explain the variability of evolutionary rates is modest compared to mRNA expression (**Figure 3**, dark blue). Notably, protein surface non-adhesiveness, the fraction of charged amino acids and effects quantified by T_m^{Agg} are likely to represent complementary sources of constraints, since they don't correlate strongly with each other (**Supplementary Table 5**).

DISCUSSION

In this work we continued the verification of the MAH using various empirical datasets, and the presented results agree with and extend our previous conclusions [8, 10]. Proteins clearly need to be stable to perform their cellular functions, and maintaining protein stability does constrain sequence evolution [19]. For example, multiple deep mutational scanning experiments demonstrate that fitness effects of substitutions correlate with their $\Delta\Delta G$, i.e. destabilizing mutations tend to be more deleterious [20-22]. Nevertheless, overall stability constraints are shared across the vast majority of proteins, and differences in protein stability do not seem to play a major role in explaining the *variability* of evolutionary rates across cellular proteins. There is no paradox here, and this conclusion is in fact consistent with multiple empirical and biophysical data beyond T_m and ΔG measurements. For example, it was demonstrated that the strength of the correlation between evolutionary rates and mRNA expression is quite similar for sites with very different contributions to protein stability, such as surface sites and sites in protein cores [18]. Moreover, increasing protein stability beyond a certain threshold is not evolutionary advantageous, and may be generally detrimental to fitness, as demonstrated by multiple examples of stability-activity tradeoffs in proteins [23-25]. If misfolding effects become harmful, for example due to a significantly increased burden of transcriptional [26] and translational [27] errors, proteins can be quickly stabilized by fixation of several substitutions [26, 27], without substantial further constraints on the rest of the protein sequence. Interestingly, across the *E. coli* and *H. sapiens* datasets we analyzed, T_m^{Agg} correlates better with protein abundances, while evolutionary rates show a stronger correlation with mRNA expression levels (**Figure 2**). This again suggests that different biological mechanisms may be driving sequence constraints related to protein aggregation and those responsible for the substantial variance of protein evolutionary rates.

Direct experimental measurements demonstrated that deleterious mutations contribute to fitness effects primarily through changes in protein function, rather than protein destabilization [21]. Analyzing long-term protein evolution, we also recently showed that functional optimality, i.e. the conservation of

protein sequence and 3-dimensional structure necessary for efficient protein function, is a substantially stronger evolutionary constraint than the requirement to simply maintain folded protein stability [28]. These results suggest that the diversity of protein evolutionary rates may be more related to protein functional effects than effects associated with protein stability.

Finally, although the feasibility of a dominant MAH contribution was indeed suggested by computational simulations [6, 7, 29], biology is an empirical science, and the fidelity of proposed hypotheses should be ultimately determined by their agreement, or lack thereof, with available experimental data. It is often possible to invoke sophisticated noise and error models to explain the absence of expected observations [9]. Nevertheless, based on the preponderance of available evidence, the MAH is unlikely to play any major role in explaining a strong negative correlation between protein abundance and evolutionary rate. Major roles are also unlikely for several other aforementioned effects, such as increased protein solubility or avoidance of non-functional interactions [10, 18]. Most importantly, the search for the main factors contributing to the substantial variability of evolutionary rates across proteins must continue.

METHODS

The protein stability data were obtained from the ProTherm database [12]. Specifically, we considered the average unfolding Gibbs free energies, ΔG, for wild type proteins measured at pH values between 4 and 9, and for temperatures between 10 and 50 degrees C.

We used the rate of non-synonymous substitutions, K_a, as a measure of protein evolutionary rate. K_a values for *Escherichia coli*, *Homo sapiens* and *Arabidopsis thaliana* were calculated using the PAML package [30] relative to *Salmonella enterica*, *Mus musculus* and *Brassica oleracea*. Orthologues were identified as bidirectional best hits in pairwise local alignments calculated with Usearch [31]. We considered for analysis only protein pairs for which corresponding alignments covered at least 70% of the shortest protein length.

Protein abundance data for all species were obtained from the whole-organism integrated datasets available in the PaxDB v.4 database [32]. mRNA expression data for *E. coli* were obtained from [33]. mRNA expression from the brain frontal cortex [34] was used for *H. sapiens*, since it was shown that protein expression in this tissue correlates best with protein evolutionary rates [6]. mRNA expression from the germinating seed was used for *A. thaliana* [35].

Protein surface non-adhesiveness was obtained from [17]. Specifically, this measure equals the negative sum of amino acid stickiness scores [17] across sequence sites located on protein surfaces based on protein 3-dimentional structures [36].

The data used for analyses in the manuscript are available in **Supplementary Tables 6-8**.

Acknowledgements

We would like to thank Drs. Purushottam Dixit and Balázs Papp for helpful scientific discussions. This work was supported by the National Institute of General Medical Sciences grant R35GM131884 to Dennis Vitkup.

Supplementary Table S1
Correlations between experimentally measured ΔG values, protein abundances, mRNA expression levels, and protein evolutionary rates

Supplementary Table S2
Correlations between genome-wide melting temperatures, protein abundances, mRNA expression levels, and protein evolutionary rates

Supplementary Table S3
Correlations between fractions of charged amino acids, protein abundances, mRNA expression levels, and protein evolutionary rates

Supplementary Table S4
Correlations between protein surface non-adhesiveness, protein abundances, mRNA expression levels, and protein evolutionary rates

Supplementary Table S5
Correlations between protein surface non-adhesiveness, fractions of charged amino acids, and melting temperatures.

Supplementary Table S6
Data on evolutionary rates, mRNA expression levels, protein abundancies, ΔG, melting temperatures, surface non-adhesiveness, and fractions of charged amino acids for *E. coli*.

Supplementary Table S7
Data on evolutionary rates, mRNA expression levels, protein abundancies, ΔG, melting temperatures, surface non-adhesiveness, and fractions of charged amino acids for *H. sapiens*.

Supplementary Table S8
Data on evolutionary rates, mRNA expression levels, protein abundancies, melting temperatures, and fractions of charged amino acids for *A. thaliana*.

References

1. Koonin, E.V., *The logic of chance : the nature and origin of biological evolution*. 2012, Upper Saddle River: PT Press.
2. Pal, C., B. Papp, and L.D. Hurst, *Highly expressed genes in yeast evolve slowly.* Genetics, 2001. **158**(2): p. 927-31.
3. Pal, C., B. Papp, and M.J. Lercher, *An integrated view of protein evolution.* Nat Rev Genet, 2006. **7**(5): p. 337-48.
4. Zhang, J. and J.R. Yang, *Determinants of the rate of protein sequence evolution.* Nat Rev Genet, 2015. **16**(7): p. 409-20.
5. Drummond, D.A., et al., *Why highly expressed proteins evolve slowly.* Proc Natl Acad Sci U S A, 2005. **102**(40): p. 14338-43.
6. Drummond, D.A. and C.O. Wilke, *Mistranslation-induced protein misfolding as a dominant constraint on coding-sequence evolution.* Cell, 2008. **134**(2): p. 341-52.
7. Yang, J.R., S.M. Zhuang, and J. Zhang, *Impact of translational error-induced and error-free misfolding on the rate of protein evolution.* Mol Syst Biol, 2010. **6**: p. 421.

8. Plata, G. and D. Vitkup, *Protein Stability and Avoidance of Toxic Misfolding Do Not Explain the Sequence Constraints of Highly Expressed Proteins.* Mol Biol Evol, 2018. **35**(3): p. 700-703.
9. Razban, R.M., *Protein Melting Temperature Cannot Fully Assess Whether Protein Folding Free Energy Underlies the Universal Abundance-Evolutionary Rate Correlation Seen in Proteins.* Mol Biol Evol, 2019. **36**(9): p. 1955-1963.
10. Plata, G., M.E. Gottesman, and D. Vitkup, *The rate of the molecular clock and the cost of gratuitous protein synthesis.* Genome Biol, 2010. **11**(9): p. R98.
11. Leuenberger, P., et al., *Cell-wide analysis of protein thermal unfolding reveals determinants of thermostability.* Science, 2017. **355**(6327).
12. Kumar, M.D., et al., *ProTherm and ProNIT: thermodynamic databases for proteins and protein-nucleic acid interactions.* Nucleic Acids Res, 2006. **34**(Database issue): p. D204-6.
13. Savitski, M.M., et al., *Tracking cancer drugs in living cells by thermal profiling of the proteome.* Science, 2014. **346**(6205): p. 1255784.
14. Becher, I., et al., *Pervasive Protein Thermal Stability Variation during the Cell Cycle.* Cell, 2018. **173**(6): p. 1495-1507 e18.
15. Mateus, A., et al., *Thermal proteome profiling in bacteria: probing protein state in vivo.* Mol Syst Biol, 2018. **14**(7): p. e8242.
16. Volkening, J.D., K.E. Stecker, and M.R. Sussman, *Proteome-wide Analysis of Protein Thermal Stability in the Model Higher Plant Arabidopsis thaliana.* Mol Cell Proteomics, 2019. **18**(2): p. 308-319.
17. Levy, E.D., S. De, and S.A. Teichmann, *Cellular crowding imposes global constraints on the chemistry and evolution of proteomes.* Proc Natl Acad Sci U S A, 2012. **109**(50): p. 20461-6.
18. Yang, J.R., et al., *Protein misinteraction avoidance causes highly expressed proteins to evolve slowly.* Proc Natl Acad Sci U S A, 2012. **109**(14): p. E831-40.
19. Dill, K. and S. Bromberg, *Molecular Driving Forces: Statistical Thermodynamics in Biology, Chemistry, Physics, and Nanoscience*. 2012: CRC Press.
20. Jacquier, H., et al., *Capturing the mutational landscape of the beta-lactamase TEM-1.* Proc Natl Acad Sci U S A, 2013. **110**(32): p. 13067-72.
21. Firnberg, E., et al., *A comprehensive, high-resolution map of a gene's fitness landscape.* Mol Biol Evol, 2014. **31**(6): p. 1581-92.
22. Sarkisyan, K.S., et al., *Local fitness landscape of the green fluorescent protein.* Nature, 2016. **533**(7603): p. 397-401.
23. Wang, X., G. Minasov, and B.K. Shoichet, *Evolution of an antibiotic resistance enzyme constrained by stability and activity trade-offs.* J Mol Biol, 2002. **320**(1): p. 85-95.
24. Tokuriki, N., et al., *How protein stability and new functions trade off.* PLoS Comput Biol, 2008. **4**(2): p. e1000002.
25. Knies, J.L., F. Cai, and D.M. Weinreich, *Enzyme Efficiency but Not Thermostability Drives Cefotaxime Resistance Evolution in TEM-1 beta-Lactamase.* Mol Biol Evol, 2017. **34**(5): p. 1040-1054.
26. Goldsmith, M. and D.S. Tawfik, *Potential role of phenotypic mutations in the evolution of protein expression and stability.* Proc Natl Acad Sci U S A, 2009. **106**(15): p. 6197-202.
27. Bratulic, S., F. Gerber, and A. Wagner, *Mistranslation drives the evolution of robustness in TEM-1 beta-lactamase.* Proc Natl Acad Sci U S A, 2015. **112**(41): p. 12758-63.
28. Konate, M.M., et al., *Molecular function limits divergent protein evolution on planetary timescales.* eLife, 2019. **8**.
29. Serohijos, A.W., Z. Rimas, and E.I. Shakhnovich, *Protein biophysics explains why highly abundant proteins evolve slowly.* Cell Rep, 2012. **2**(2): p. 249-56.

30. Yang, Z., *PAML: a program package for phylogenetic analysis by maximum likelihood.* Comput Appl Biosci, 1997. **13**(5): p. 555-6.
31. Edgar, R.C., *Search and clustering orders of magnitude faster than BLAST.* Bioinformatics, 2010. **26**(19): p. 2460-1.
32. Wang, M., et al., *Version 4.0 of PaxDb: Protein abundance data, integrated across model organisms, tissues, and cell-lines.* Proteomics, 2015. **15**(18): p. 3163-8.
33. McClure, R., et al., *Computational analysis of bacterial RNA-Seq data.* Nucleic Acids Res, 2013. **41**(14): p. e140.
34. Mele, M., et al., *Human genomics. The human transcriptome across tissues and individuals.* Science, 2015. **348**(6235): p. 660-5.
35. Klepikova, A.V., et al., *A high resolution map of the Arabidopsis thaliana developmental transcriptome based on RNA-seq profiling.* Plant J, 2016. **88**(6): p. 1058-1070.
36. Levy, E.D., *A simple definition of structural regions in proteins and its use in analyzing interface evolution.* J Mol Biol, 2010. **403**(4): p. 660-70.